基于多尺度特征预测的异常事件检测

王 军*

(电子科技大学中山学院机电工程学院 广东 中山 528402)

【摘要】传统异常事件检测方法面临着视频中物体大小变化、背景等问题的影响。为了解决该问题，提出了一种基于多尺度特征预测的异常事件检测方法。首先，利用空洞卷积提取不同大小感受野的特征并进行融合以解决物体大小变化的问题。然后，使用一种轻量化的通道注意力方法来减少无效背景信息的影响。最后，为了充分利用视频帧之间的上下文信息，采用深度特征预测模块根据历史时刻的特征预测当前时刻的特征，并根据预测特征与真实特征之间的差异进行异常判断。在 USCD Ped2，UMN 两个基准数据集上进行了实验，实验结果表明了该文方法的有效性。

关 键 词 异常事件检测； 通道注意力； 特征预测； 多尺度特征

中图分类号 TP391 **文献标志码** A doi:10.12178/1001-0548.2021333

Abnormal Event Detection Based on Multi-Scale Features Prediction

WANG Jun*

(College of Mechanical and Electrical Engineering, Zhongshan Institute, University of Electronics Science and Technology Zhongshan Guangdong 528402)

Abstract A novel method for abnormal event detection is proposed based on multi-scale feature prediction. Firstly, dilated convolution network is used to extract the features of different size receptive fields and fuse them so that address the objects of different scale in video frame. Secondly, a lightweight channel-wise attention module is applied to reduce the impact of background information. Finally, in order to make full use of the context information between video frames, a deep feature prediction module is applied to predict the features of the current moment based on the features of the historical moment, and the prediction error is used for abnormality judgment. Experiments were performed on the two benchmark data sets of USCD Ped2 and UMN to test and evaluate the proposed method. The experiments results show that the proposed method is more effective than other state-of-the-art methods.

Key words abnormal event detection; channel-wise attention; feature prediction; multi-scale feature

随着公共安全体系建设的不断发展，监控摄像头被广泛应用在各种公共场合中，如商场、街道、银行等。由于监控视频内容庞大，人工进行异常事件检测会耗费大量的人力物力[1-4]。因此，如何建立一个高效的自动异常事件检测系统非常重要，这也是计算机视觉研究的一个重要方向。

异常事件检测大体可分为基于手工特征的方法和基于深度学习的方法，近年来基于深度学习的方法被广泛研究[1,5-10]。由于深度神经网络卓越的生成能力，基于重建和预测的异常事件检测方法被广泛地使用。文献 [1] 开创性地将 U-net 网络引入异常事件检测领域中，根据历史时刻的视频帧预测未来帧，并根据预测误差进行异常检测。文献 [5] 对 U-Net 网络进行改进，将其变化为一个双流网络，网络的两个流分别对视频帧进行重建和预测，并引入生成对抗的思想进行训练，以生成更加逼真的图像，最后根据重建误差进行异常判断。考虑到视频是由一系列关联性很强的图像组成，不少学者提出时间信息的概念，并将其用于视频异常事件检测中。文献 [7] 利用 3D 卷积提取输入视频片段中的空间特征和时间信息特征，并使用两个 3D 反卷积分别进行重建和预测。循环神经网络 (recurrent neural network, RNN) 及其变体由于其优秀的时间信息编码能力被用于异常事件检测中。文献 [8] 将

收稿日期：2021－11－10；修回日期：2022－04－11
基金项目：国家自然科学基金 (51678075)
作者简介：王军 (1971－)，男，博士，副教授，主要从事图像处理与模式识别方面的研究.
*通信作者：王军，E-mail：106919257@qq.com

LSTM 网络与软硬注意力相结合提出行人轨迹预测网络，该网络不仅关注行人的历史轨迹，同时还关注该行人的邻域对其轨迹的影响。文献 [9] 将卷积自编码器与 ConvLSTM 相结合，利用卷积自编码器获取空间特征的变化，利用 ConvLSTM 记录特征随时间的变化，并将光流作为补充信息，从全局−局部的角度分析异常。此外，由于监控视频的视角大多是固定的，视频中可能会出现不同大小的物体，因此多尺度特征被引入到检测模型中。文献 [10] 提出一种双边多尺度聚合网络，该网络利用不同膨胀率的空洞卷积提取不同大小感受野的特征，利用 ConvLSTM 进行双边时间信息编码。

虽然视频异常检测已经取得了一些成就，但依然存在一些问题。如视频中物体大小的变化、复杂背景的影响以及不同场景下异常的定义不同等。为了解决以上问题，本文提出一种充分利用多尺度特征和时间−空间信息的异常事件检测方法。首先，利用经过预训练的 VGG16 网络提取特征，构建多尺度特征融合模块获取更多不同大小感受野的信息，以获得对输入视频帧的完备表示。其次，使用一种轻量化的通道注意力模块来强调视频中重要的前景信息，以减少背景信息对检测的影响。在此基础上，根据历史时刻特征预测当前时刻的特征，这将有助于弥补前文模块中对上下文信息和时间信息利用不足的缺陷。在训练阶段，最小化预测特征与真实特征之间的欧式距离使整个网络收敛。在测试阶段，本文认为仅包含正常事件的视频帧可以很好地预测，而包含异常事件的视频帧将会产生很大的预测误差。因此，在测试时将根据预测误差进行异常判断。在 USCD Ped2 和 UMN 两个基准数据集上进行了实验，实验结果表明了提出方法的有效性。

1　基于空洞卷积的多尺度特征提取

为了编码尽可能多的空间信息，使用空洞卷积网络构建一种多尺度特征融合模块，以获得包含输入视频帧的全局−局部信息的特征图。

由于视频帧中存在不同大小的对象，所以不同大小感受野的信息在异常事件检测中非常重要，而空洞卷积[11]可以通过调整膨胀率来获得不同大小感受野的特征语义，因此本文利用空洞卷积设计了一种具有多分支结构的多尺度特征融合模块，用于提取视频的多尺度特征，其结构如图 1 所示。输入的视频帧首先经过一个预训练的 VGG16网络进行特征提取，取 VGG16 第三个池化层的输出作为多分支结构的输入。输入的特征图被送入 4 个不同的分支中进行处理。第一个分支用于保留原始特征信息，其余 3 个分支通过具有不同膨胀率的空洞卷积提取多尺度特征。第二、第三、第四分支的膨胀率分别为 1、3、5，则其卷积核对应的感受野分别为 3×3、7×7、11×11。由于空洞卷积的存在，可以在不做池化损失信息的情况下，增大特征图的感受野，让卷积的输出包含较丰富的信息。

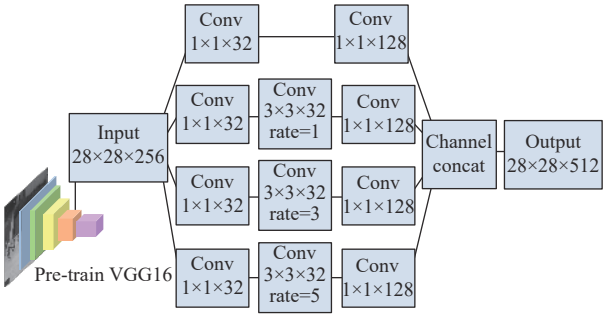

图 1　多尺度特征融合模块结构

在 4 个分支中，小膨胀率的卷积核有利于提取视频帧中小物体的特征，而大膨胀率的卷积核有助于提取视频帧中大物体的特征。本文在空洞卷积的前后增加了 1×1 的卷积来调整特征图的通道数，以减少模型的参数量和运算量。最后，将 4 个分支的特征图在通道上进行拼接，获得一个包含全局−局部信息的特征表示U。

2　基于通道注意力的背景抑制

在视频异常事件检测中，监控摄像头通常是固定的，因此画面中可能存在大量静止的区域。异常事件通常发生在运动变化的前景物体上，因此希望网络能够重点关注运动变化的前景物体。在特征图中，不同的通道包含着不同的语义信息，有的通道包含着静止的背景信息，有的通道包含着变化的前景信息。为了减少背景信息对检测的影响，强调当前帧中重要的前景物体的信息，本文引入通道注意力机制。通道注意力通过计算各个通道中包含的信息以及通道之间的关系生成通道的权重，并将权重赋予其对应的通道。本文基于 SENet 设计了一种轻量化通道注意力模型，包含挤压、激活、重新分配权重 3 个步骤，其结构如图 2所示。其中，挤压 (squeeze) 是通过在输入特征图的每个通道上执行全局平均池化得到特征图的全局压缩特征向量；激活 (excitation) 是通过两组 1×1 卷积、批正则化、

激活函数获得输入特征图中每个通道的权值；重新分配权重操作 (reassign weights) 是将权值对输入的特征进行加权。

首先在挤压操作中，输入特征图U经过全局平均池化从$H\times W$的大小池化成一个一维向量，该过程可表示为：

$$z_c = F_{\mathrm{sq}}(U_c) = \frac{1}{WH}\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i,j) \tag{1}$$

式中，$U_c(i,j)$表示输入特征U的第c个通道(i,j)位置上的空间信息；$F_{\mathrm{sq}}(U_c)$表示挤压操作；z_c表示空间描述符。

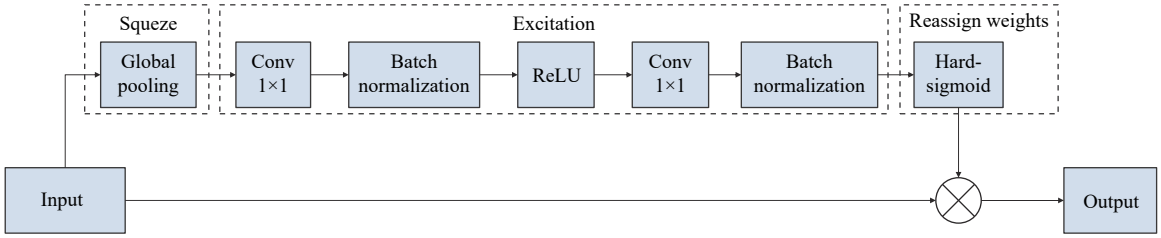

图 2　轻量化通道模块注意力结构

挤压操作之后是激活操作，现有的通道注意力机制通常使用全连接层来计算通道之间的关系和权重，这无疑会增加运算的复杂度并且可能会导致过拟合。本文使用两个 1×1 的卷积来替换全连接层，以减少运算量。在每一个卷积层之后使用批正则化层进行正则化，以重新调整数据的分布，保证训练过程中梯度的有效性。在两个批正则化后面，分别使用 ReLU 和 Hard-sigmoid 函数作为激活函数。激活操作可表示为：

$$\boldsymbol{S} = F_{\mathrm{ex}}(z) = \sigma(N(\boldsymbol{W}_2 N(\delta(\boldsymbol{W}_1 z)))) \tag{2}$$

式中，z表示经过挤压后得到的一维向量；$\boldsymbol{W}_1$和$\boldsymbol{W}_2$分别表示两个卷积层的权重；N表示批正则化；σ和δ分别表示 hard-sigmoid 激活函数和 ReLU 激活函数；$F_{\mathrm{ex}}(z)$表示激活操作；$\boldsymbol{S}$表示通道注意力权值，为一维向量，维度等于输入特征图的通道数 512。权值中某个维度的值越高，表明其对应的通道的重要性越高。

最后，在重新分配权重中，将输入特征图U与通道权重相乘，强调输入特征图中重要的通道信息。

重新调整通道权重：

$$F_{\mathrm{att}} = F_{\mathrm{scale}}(U,\boldsymbol{S}) = S_c U_c \tag{3}$$

式中，S_c表示第c个通道的注意力权重；U_c表示输入的第c个通道的特征图；$F_{\mathrm{scale}}(U,\boldsymbol{S})$表示重新调整权重操作；$F_{\mathrm{att}}$表示进行注意力计算后的通道注意力特征图。

3　基于特征预测的异常事件检测

正常事件状态变化比较平稳，可以预测，而异常事件状态通常会出现突变，不可预测。因此可以通过比较某帧的预测特征和真实特征来判断事件是否异常。

监控视频是由一系列关联性很强的视频帧组成，为了充分利用视频帧之间的时间信息，本文构建了深度特征预测模块。该模块根据历史时刻的特征图预测当前时刻的特征图。将经过注意力模块后获得的连续 5 个历史时刻的特征图在通道上进行拼接，组成深度特征预测模块的输入$\boldsymbol{X}^t$。由于输入的特征图通道数较高，因此本文设计了一个仅包含 1×1 卷积核、ReLU 激活函数的特征预测模块，该模块由编码器、解码器组成，其具体结构如表 1 所示。

表 1　特征预测模块的结构

Layer	Filter/Stride	Activation function
Conv1	(1×1×512)/1	ReLU
Conv2	(1×1×256)/1	ReLU
Conv3	(1×1×128)/1	–
Conv4	(1×1×256)/1	ReLU

在深度特征预测模块中，编码器计算不同时刻特征图之间的关系，并将其映射到一个低维空间中，解码器根据低维空间中的特征预测当前时刻的特征图。预测特征图与真实特征图之间的差异将被用于异常判断。

训练时，在仅包含正常数据样本的训练集中对网络进行训练，最小化预测特征与真实特征之间的欧式距离来对整个网络进行训练：

$$L = \left\| F_{\mathrm{Pred}}^{t} - \phi^{t} \right\|_2 \tag{4}$$

式中，F_{Pred}^{t}表示当前时刻预测特征图；ϕ^t表示当前

时刻 VGG16 第三个池化层输出特征图。

在测试时，根据当前时刻 VGG16 提取的特征图与预测特征图之间的差异来进行异常判断，计算预测特征图与 VGG16 第三个池化层输出特征图之间的欧式距离，若误差大于设定的阈值α，则说明输入的视频片段中存在异常。公式为：

$$s_t = \left\| F_{\text{Pred}}^t - \phi' \right\|_2 \tag{5}$$

式中，s_t表示测试时当前时刻的异常得分。

4　实验与结果

在两个公开数据集 UCSD Ped2[19] 和 UMN[14] 上验证本文方法的有效性。它们的训练数据都仅包含正常样本。

4.1　实验数据

UCSD 数据集通过学校里固定在较高位置上俯瞰人行道的摄像机获得，本文仅使用 Ped2 进行实验。Ped2 中含有骑自行车、滑旱冰、小汽车等异常事件，共有 16 个训练视频样本和 12 个测试视频样本。

UMN 数据集包含 3 个不同的场景和 11 个视频片段，训练集包含 3300 帧，测试集包含 4439 帧。其异常事件主要包括人群单方面跑动、人群四散等。

4.2　实验设置

使用的深度学习训练框架为 Pytorch，所有的实验都基于 NVIDIA RTX2080Ti。将输入的视频帧大小调整到 224×224 以满足 vgg16 的输入标准。训练时使用随机梯度下降法进行参数优化，学习率设置为 1×10^{-4}，并在训练 100 轮后将其降低至 1×10^{-5}。选取帧级别的 ROC 曲线及 ROC 曲线下面积 AUC 作为异常行为的评价指标，在该评估方法中，只要当前帧中存在异常特征，则立即判断该视频帧为异常帧。

4.3　消融实验

4.3.1　多尺度特征的影响

为了证明多尺度特征融合的有效性，在基线网络 U-Net[1] 的瓶颈层中添加多尺度模块来进行消融实验。

实验中修改 U-Net 的输入为单个视频帧，输入视频帧经过一系列的卷积层进行特征提取，利用反卷积和跳转连接进行图像重建。计算重建图像与输入图像的欧式距离来判断输入视频帧是否存在异常。在评价指标上，对比了平均正常得分和平均异常得分之差Δ_s，Δ_s的值越大，模型对正常事件和异常事件的区分能力越强，从而说明特征在异常事件检测中的可分性越好。实验结果如表 2 所示，与基线网络相比，使用多尺度特征融合后平均正常得分与平均异常得分的差值更大，这说明在 U-Net 的瓶颈层添加的多尺度模块编码了更多的空间特征，解码器可以利用更多的特征来对图像进行重建。因此添加了多尺度模块的基线网络可以获得更好的效果。

表 2　不同方法在 USCD Ped2 和 UMN 上的Δ_s对比结果

方法	数据集	
	Ped2	UMN
U-Net	0.435	0.362
U-Net with Multi-scale feature fusion module	0.468	0.395

4.3.2　通道注意力的影响

为了证明所提出的通道注意力的有效性，本文在结合了多尺度特征的基线网络上进行通道注意力的实验。实验首先在基线网络 U-Net 上添加多尺度特征融合模块，其次在多尺度特征融合模块后面添加通道注意力进行对比实验。与前一节的实验评价方法一样，对比平均正常得分与平均异常得分之间的差值。

实验结果如表 3 所示。由实验结果可知，在不使用注意力的情况下，网络对特征图中的所有通道同等看待，容易受到浅层特征中噪声以及背景等因素的干扰，因此获得的检测效果较差。而在多尺度特征融合模块后面添加注意力机制后，正常得分与异常得分之间的差值变大，这表明通道注意力可以有效地减少背景冗余信息，增加运动变化的前景物体信息在特征图中的权重。此外，在 SENet 中使用多层感知机来计算不同通道间的关系来获取各个通道的权重，这不可避免地容易造成过拟合，使得检测效果下降，而本文利用两个 1×1 的卷积来替换多层感知机，并在其后面添加批正则化来保证训练过程中梯度的有效性，避免了过拟合的现象，同时减少了模型的参数量，因此获得的实验结果较好。

表 3　不同通道注意力的对比实验结果

方法	数据集	
	Ped2	UMN
Without channel-wise attention	0.468	0.395
With SENet	0.493	0.413
With proposed attention module	0.502	0.429

4.4 对比实验

在帧级别的评估方法下，将提出的方法与已有方法在 Ped2 数据集上进行对比，其中包括基于手工特征的方法以及基于深度学习的方法[13-18]。在异常检测中，将异常判断的阈值设置为 0.1、0.2、0.3、0.5、0.8，可以计算出 5 组不同的假阳率 (false positive rate, FPR) 和真阳率 (true positive rate, TPR)。以 FPR 为横坐标，TPR 为纵坐标，绘制出 ROC 曲线，ROC 曲线下的面积即为 AUC，面积越大，则检测的效果越好。

在 Ped2 数据集上的实验结果 (ROC 曲线) 如图 3 所示，本文方法在帧级别下，获得了最好的效果。Social Fore[14] 仅使用了手工特征的方法，因此其得到的帧级别 AUC 仅为 0.556。文献 [13] 将外观特征和运动特征结合起来，帧级别下 AUC 提升至 0.850，但其遗漏了时间信息。其他方法如 Unmasking[15]、Hashing[16]、spatiotemporal saliency detector[18] 以及文献 [19] 使用使用 MDT(mixtures of dynamic textures) 方法在帧级别下 AUC 分别获得了 0.822、0.910、0.877、0.875 的检测效果。以上方法的数据均来自于文献原文。本文方法由于考虑了视频中全局−局部特征，并充分利用了时间信息，因此获得了更好的检测效果，帧级别下 AUC 达到了 0.925。

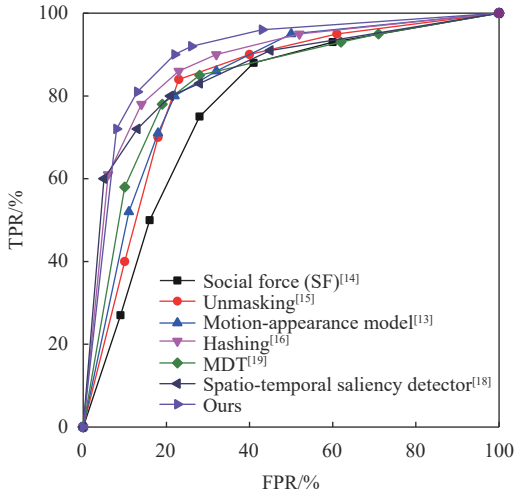

图 3 不同方法在 Ped2 上的帧级别 ROC 曲线对比

在 UMN 数据集上验证本文方法，并将所得的结果与上面的方法进行帧级别 ROC 比较，结果 (ROC 曲线) 如图 4 所示。本文方法同样获得了最好的结果，帧级别 AUC 达到了 0.991。基于手工特征的方法 Social Fore[14] 获得的帧级别 AUC 为 0.96，将外观特征与运动特征相结合的方法 Motion-appearance model[13] 获得的帧级别 AUC 为 0.983；将 Unmasking 迁移至异常事件检测的方法[15] 获得的帧级别 AUC 为 0.951；基于 Hashing filter[16] 的方法，基于 spatiotemporal saliency detector[18] 的方法以及文献 [19] 的方法分别获得的帧级别 AUC 为 0.987、0.938、0.961。

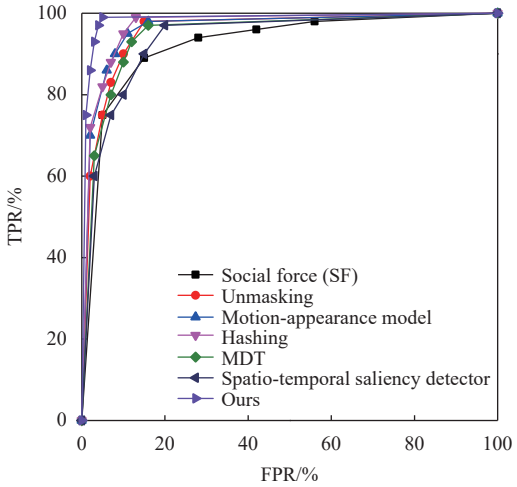

图 4 不同方法在 UMN 上的帧级别 ROC 曲线对比

5 结 束 语

本文提出了一种充分利用视频中多尺度信息和时间信息的异常事件检测网络，该网络不仅关注视频中的全局−局部信息，还考虑了空间−时间信息。该网络利用空洞卷积获取多个不同大小的感受野的信息并进行融合以获得整个视频帧的全局−局部表示，并且引入一种轻量化通道注意力机制，通过计算特征图中不同通道所含信息的重要程度，提升重要通道的权重，抑制背景和噪声等干扰因素的影响。最后，为了充分利用时间信息，使用自编码器编码历史时刻的特征序列并预测当前时刻的特征，预测特征与真实特征之间的误差将被用于异常判断。在两个基准数据集上与几种方法进行了对比实验，实验结果证明了本文方法的有效性。

参 考 文 献

[1] LIU W, LUO W, LIAN D, et al. Future frame prediction for anomaly detection-a new baseline[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 6536-6545.

[2] LU C, SHI J, JIA J. Abnormal event detection at 150 fps in matlab[C]//2013 IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2013: 2720-2727.

[3] SULTANI W, CHEN C, SHAH M. Real-World anomaly

detection in surveillance videos[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 6479-6488.

[4] SONG H, SUN C, WU X, et al. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos[J]. IEEE Transactions on Multimedia, 2020, 22(8): 2138-2148.

[5] NGUYEN T N, MEUNIER J. Anomaly detection in video sequence with appearance-motion correspondence[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). [S.l.]: IEEE, 2019: 1273-1283.

[6] CHANG Y, TU Z, XIE W, et al. Clustering driven deep autoencoder for video anomaly detection[C]//16th European Conference on Computer Vision (ECCV). Glasgow: Springer, 2020: 329-345.

[7] ZHAO Y, DENG B, SHEN C, et al. Spatio-temporal autoencoder for video anomaly detection[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: Association for Computing Machinery, 2017: 1933-1941.

[8] FERNANDO T, DENMAN S, SRIDHARAN S, et al. Soft + hardwired attention: An lstm framework for human trajectory predictionand abnormal event detection[J]. Neural Networks 2018, 108: 466-478.

[9] YANG B, CAO J, WANG N, et al. Anomalous behaviors detection in moving crowds based on a weighted convolutional autoencoder-long short-term memory network[J]. IEEE Transactions on Cognitive and Devel-Opmental Systems, 2019, 11(4): 473-482.

[10] LEE S, KIM H G, RO Y M. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection[J]. IEEE Transactions on Image Processing, 2020, 29: 2395-2408.

[11] ABATI D, PORRELLO A, CALDERARA S, et al. Latent space autoregression for novelty detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2019: 481-490.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018, 7132-7141.

[13] ZHANG Y, LU H C, ZHANG L H, et al. Combining motion and appearance cues for anomaly detection[J]. Pattern Recognition, 2016, 51: 443-452.

[14] MEHRAN R, OYAMA A, SHAH M. Abnormal crowd behavior detection using social force model[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2009: 935-942.

[15] CONG Y, YUAN J, LIU J. Sparse reconstruction cost for abnormal event detection[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado: IEEE, 2011: 3449-3456.

[16] ZHANG Y, LU H C, ZHANG L H, et al. Video anomaly detection based on locality sensitive hashing filters[J]. Pattern Recognition, 2016, 59: 302-311.

[17] LIU Y S, LI C L, P´oczos Barnaba´s. Classifier two-sample test for video anomaly detections[EB/OL]. [2021-10-11]. http://www.bmva.org/bmvc/2018/contents/papers/0237.pdf.

[18] WANG Y, ZHANG Q, LI B. Efficient unsupervised abnormal crowd activity detection based on a spatiotemporal saliency detector[C]//2016 IEEE Winter Conference on Applications of Computer Vision (WACV). [S.l.]: IEEE, 2016: 1-9.

[19] MAHADEVAN V, LI W, BHALODIA V, et al. Anomaly detection in crowded scenes[C]//Computer Vision & Pattern Recognition. [S.l.]: IEEE, 2010: 1975-1981.

编　辑　叶　芳